R E P O R T R E S U M E S

ED 014 328 PS 000 220

SUMMARY REPORT OF A STUDY OF THE FULL-YEAR 1966 HEAD START PROGRAMS.

PLANNING RESEARCH CORP., WASHINGTON. D.C.
REPORT NUMBER PRC-R-1042 PUB DATE 22 SEP 67
REPORT NUMBER OEO-1308
EDRS PRICE MF-$0.25 HC-$1.16 27P.

DESCRIPTORS- TEST INTERPRETATION, *DISADVANTAGED YOUTH, *RATING SCALES, BEHAVIOR DEVELOPMENT, MATURITY TESTS, EDUCATION OBJECTIVES, *EDUCATIONAL TESTING, POST TESTING, COMPARATIVE TESTING, *TEST RESULTS, *PROGRAM EFFECTIVENESS, CHILD DEVELOPMENT, HEADSTART, PPVT, PSI, BI, VSMS, DAP

THIS SUMMARY OF SELECTED HIGHLIGHTS IS FROM A MAJOR REPORT TITLED "A STUDY OF THE FULL-YEAR 1966 HEAD START PROGRAMS." THE STUDY WAS DONE TO DETERMINE WHETHER THE PERFORMANCE OF CHILDREN ON FIVE TESTS AND RATING SCALES IS RELATED TO THE LENGTH OF THE 1966 FULL-YEAR PROGRAM WHICH THEY ATTENDED. FULL-YEAR PROGRAMS WERE CLASSED AS SHORT TERM FOR 15 WEEKS, OR LESS, MEDIUM TERM FOR 17 TO 23 WEEKS, AND LONG TERM FOR 25 WEEKS OR MORE. NINETEEN TESTERS WHO FULFILLED SPECIAL REQUIREMENTS WERE CHOSEN TO ADMINISTER THE TESTS. IN ALL, 964 CHILDREN IN 72 CENTERS WERE TESTED. TESTS AND SCALES USED AND BRIEFLY DISCUSSED WERE THE PEABODY PICTURE VOCABULARY TEST, THE REVISED PRE-SCHOOL INVENTORY, THE BEHAVIOR INVENTORY, THE VINELAND SOCIAL MATURITY SCALE, AND THE DRAW-A-PERSON TEST. RECOMMENDATIONS ARE MADE REGARDING THE USE OF THESE TESTS. FROM ANALYSIS OF TEST SCORES IT WAS DETERMINED THAT THERE WAS NO RELIABLE EVIDENCE OF AN AVERAGE DIFFERENCE IN PERFORMANCE RELATED TO LENGTH OF PROGRAM ATTENDANCE. SOME UNRESOLVED QUESTIONS RAISED BY THE STUDY ARE GIVEN. EVIDENCE INDICATES A NEED FOR THE SPELLING OUT OF SPECIFIC GOALS AND OBJECTIVES FOR HEAD START PROGRAMS. (INCLUDES A COMMENTARY ON THIS REPORT BY JOHN MCDAVID.) (EF)

ED014328

SUMMARY REPORT OF A STUDY OF THE FULL-YEAR 1966 HEAD START PROGRAMS

PRC R-1042

22 September 1967

Prepared for

Office of Economic Opportunity
Project Head Start
Division of Research and Evaluation

PS000220

PLANNING RESEARCH CORPORATION
LOS ANGELES, CALIFORNIA WASHINGTON, D. C.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

SUMMARY REPORT OF A STUDY OF THE FULL-YEAR 1966 HEAD START PROGRAMS

PRC R-1042

22 September 1967

Prepared for

Office of Economic Opportunity
Project Head Start
Division of Research and Evaluation
Under Contract OEO-1308

PLANNING RESEARCH CORPORATION

Universal Building North • Suite 1030 • 1875 Connecticut Avenue N.W. • Washington, D.C. 20009

PS000220

TABLE OF CONTENTS

		Page
A.	Tests	3
B.	Untested and Untestable Children	6
C.	Principal Results	7
D.	Conclusions	17
E.	Commentary: OEO - Head Start Office of Research and Evaluation	21

LIST OF EXHIBITS

		Page
1.	Unadjusted and Adjusted Raw Score Means	8
2.	Characteristics of Short-, Medium-, and Long-Term Eligible Children	11

The Planning Research Corporation (PRC) conducted a study of the relationship of the performance of children on certain tests or rating scales to the length of the 1966 Full-Year Head Start programs which they attended. From analyses of the test scores obtained at the end of samples of programs, it was concluded that there was no reliable evidence of an average difference in performance related to length of program. Program lengths studied covered a range of 6 to 36 weeks. Thus, within the limits of the study design (including the programs sampled), instruments, measures, and methods of analysis employed, the study was not able to demonstrate an effect of Head Start--either positive or negative--related to the length of the programs.

A report, A Study of the Full-Year 1966 Head Start Programs, was submitted to the Office of Economic Opportunity (OEO) on 31 July 1967, and contains a detailed account of the procedures and findings of the study. This summary presents selected highlights from the more comprehensive report and provides appropriate references to more detailed and technical explanations within the major report. Page references in parentheses refer the reader to related pages in the complete Final Report.[1]

The study was undertaken as part of a national evaluation of the 635 1966 full-year Head Start programs throughout the country. At the time the study was initiated, the U.S. Bureau of the Census was already obtaining information from a national sample of Child Development Centers (CDC's) or Head Start centers, children, and parents. These data included:

- Age, sex, and race characteristics of the children.
- Medical and dental history and status of the children.
- Characteristics of staff members and workers in the sample CDC's.

---

[1] Planning Research Corporation, PRC R-886, A Study of the Full-Year 1966 Head Start Programs, H. Russell Cort, Jr., William D. Commins, Jr., Naomi H. Henderson, Margaret M. Mattis, Ruth Ann O'Keefe, and Reginald C. Orem, July 1967 (Unclassified) (Report prepared for the Office of Economic Opportunity under Contract No. OEO-1308)

- Evaluation of the individual programs by staff members and workers.
- Participation of parents in the local programs.
- Structure and other socioeconomic characteristics of the families of the children.

It was one of PRC's tasks to obtain direct measurements of the children themselves.

Unlike the summer Head Start programs which ran for approximately 8 weeks, the full-year programs, during their first year, varied in starting times, ending times, and length. Some programs had commenced in the fall of 1965, while others were just beginning in May 1966 when this study was initiated. Some programs ended in March or April 1966; others were not scheduled to end until January or February of 1967. The length of the programs varied from 40 weeks for some to less than 10 weeks for others.

This diversity offered an opportunity for evaluation of the programs that was not possible with the short, fixed-length summer programs. The intent of the study was to compare programs of different durations by determining the relationship between length of program and level of performance, achievement, or behavior of participating Head Start children. The general approach was to administer tests to sample groups of children in programs of different durations, and to obtain ratings of behaviors and abilities of these children from their teachers. The principal independent variable was the length of the individual programs, and the dependent variables were the performances or ratings of the children as measured by scores on the tests or scales.

Three main program durations were defined: short term, medium term, and long term. Short-term centers operated 15 weeks or less, medium-term centers lasted from 17 to 23 weeks, and long-term centers continued for 25 weeks or more. The fact that a program was in operation a certain number of weeks did not mean, however, that a particular child attended the program for that many weeks. While there was generally a strong correlation between the length of a program and the length of time a particular child was in the program, the two variables were

not synomymous. Consequently, two main types of analyses, based on the separate variables of program duration and program exposure, had to be made. There were essentially two major samples of children dealt with in the analyses: one sample contained 964 children, the total number of children tested, and the other sample contained 831 children, the number of "eligible" children (i.e., children within the total sample actually in attendance in their Head Start programs for the short-, medium-, or long-term duration).

The major methodological constraint on the study design resulted from the timing of the study. Since the study was initiated in May 1966, pretesting of children was impossible. Consequently, the overall approach was based on only one testing, an end-of-program testing, for all children.

A. Tests

The primary instruments used in the testing program were the Peabody Picture Vocabulary Test (PPVT) Form B, the revised Pre-School Inventory (PSI), the Operation Head Start Behavior Inventory (BI), and the Vineland Social Maturity Scale (VSMS). A Draw-A-Person (DAP) test was also administered.

The Peabody Picture Vocabulary Test is an individual test of verbal ability which does not require a verbal response. For example, the tester shows a child a page containing four pictures and says, "Show me 'table.'" There are 150 possible pictures which the child may identify. The "raw score" is essentially the number of correctly identified pictures and can be converted into mental age or intelligence quotient equivalents if desired. Form B of the test was used throughout this study.[1]

The revised Pre-School Inventory is an individually administered 85-item test of school readiness developed for Head Start by Dr. Bettye Caldwell and Mr. Donald Soule of the New York State University at Syracuse. The test provides a total score, as well as four subtest scores of separate factors related to school readiness: personal-social

---

[1] Dunn, Lloyd M., Expanded Manual, Peabody Picture Vocabulary Test. Minneapolis, Minnesota: American Guidance Service, Inc., 1965

responsiveness, associative vocabulary, numerical concept activation, and sensory concept activation. The tester administers all items to the child; the highest possible score is 90, as five items have a score value of 2 points, while all other items have a score value of 1 point. The total raw score can be broken down into scores for each of the four factors or subtests.

The Behavior Inventory is an instrument developed by Dr. Edward Zigler for the 1965 summer Head Start program. It is a set of 50 rating scales which are intended to obtain information on a number of behavioral characteristics of children. Twenty-five of the scales are intended to tap positive behavioral characteristics (such as "Is usually carefree; rarely becomes apprehensive or frightened,"), and 25 are intended to tap negative characteristics (such as "Has little respect for the rights of other children; refuses to wait his turn; usurps toys other children are playing with," etc.). The teacher, or someone else who knows the child well, rates each child in her class on each of the 50 items or scales. The BI can provide an overall behavior-adjustment score and/or a separate adjustment score for each of nine behavioral categories (sociability-cooperation-politeness; independence-dependence; curiosity-enthusiasm-exploration-creativity; persistence; emotionality; self-confidence; jealousy-attention-seeking; achievement; and leadership).

The Vineland Social Maturity Scale is an interview schedule which is given to someone who knows the child well. It provides an indication of the child's social development, maturity, and independence. First developed by Dr. Edgar A. Doll in 1935, the Vineland has a 1965 edition that was used throughout this testing program. The scale attempts to evaluate the child in eight different areas: self-help, general; self-help, eating; self-help, dressing; locomotion; occupation; communication; self-direction; and socialization.[1] It provides a total score and conversions to Social Age (SA) and Social Quotient (SQ) can be made.

---

[1] Doll, Edgar A., Vineland Social Maturity Scale, Condensed Manual for Directions, 1965 Edition. Minneapolis, Minnesota: American Guidance Service, Inc., 1965

The Draw-A-Person test (or Draw-A-Man, as originally conceived by Dr. Florence Goodenough in 1926) is a simple test of intelligence in which the tester asks the child to draw a person. The drawing is later scored against a detailed list of criteria, and total points for each drawing are converted into standard scores (IQ's) or mental age equivalents. In this study the children's drawings were scored according to the criteria delineated by Harris[1] for male and female drawings.

In addition, information was used from three Head Start forms collected by the Bureau of the Census: the Staff Members Information Form, the Paid and Volunteer Workers Evaluation Form, and the Family Information Form.

The procedure for selecting children (p. 10)[2] was designed so that each child in each of the 72 short- (15 weeks or less), medium- (17 to 23 weeks), or long-term (25 weeks or more) centers that comprised the sample had the same probability of being included in the sample of children. Testers were given explicit procedures for selecting children, and in no case known did testers select or test children on any basis other than the sampling procedures given them (p. 5).

The following requirements were emphasized in selecting the 19 testers from more than 200 candidates:

- A college degree in a field related to education, sociology, psychology, or guidance.
- Experience with preschoolers or disadvantaged children.
- Language fluency in Spanish as well as English.
- A flexible summer schedule.

All testers met at least three of the four requirements, and each tester submitted a recommendation from one of his college instructors (pp. 13-15 and Appendix C).

---

[1] Harris, Dale B., Children's Drawings as Measures of Intellectual Maturity. New York: Harcourt, Brace, and World, 1963

[2] Throughout this summary, references in parentheses refer to the pages in the Final Report, A Study of the Full-Year 1966 Head Start Programs, where a more extensive discussion can be found.

Two tester-training sessions were held at the Center of Adult Education at the University of Maryland, with each session lasting approximately 3-1/2 days and covering the following areas:

- Orientation to the project and to test administration.
- Tester-child relationships with emphasis on the culturally deprived child.
- Adult-adult relationships.
- Role-playing in possible test situations.
- Procedures in data collection and scoring.
- Practice in testing young children.
- Practice in interviewing.

Subsequent to training, each tester was observed in the field at least once by a PRC supervisor.

There were enough Spanish-speaking children in the sample (67 eligible children) to warrant special testing procedures. Testers fluent in Spanish were sent to areas known to have many children for whom Spanish was the primary language. When a tester encountered a Spanish-speaking child, he administered the tests in Spanish and made appropriate notations on the child's test data. Data for these children were analyzed separately from the rest of the samples.

In summary, 964 children in 72 1966 full-year Head Start programs throughout the country were tested between May and August 1966. Between 12 and 15 children were tested in each of the 72 centers. Of the 964 children tested, only 831 were actually participants in their programs for approximately the same length of time their programs were in operation. These 831 children were labeled "eligible." The remaining 133 children were considered ineligible for certain analyses because their actual attendance in Head Start was considerably less than their program's term of operation. Five instruments--the Peabody Picture Vocabulary Test, Pre-School Inventory, Draw-A-Person, Vineland Social Maturity Scale, and Behavior Inventory--were administered for each child.

B. Untested and Untestable Children

Some children in the original sample could not be tested. In some cases, the child was simply unavailable due to absence; in other cases,

the child was untestable because he would not respond to the tester in the test situation. Testers generally spent from 20 to 60 minutes attempting to gain a child's cooperation before deciding to consider the child untestable. Absent children included children who had never attended the program but had been registered, children who were ill, children who had been withdrawn from the program for a variety of reasons, and children whose families had moved. Untestable children included children who (1) refused to go with the tester to the testing room, (2) were extremely reticent, (3) were tearful and uncommunicative, (4) spoke unintelligibly, and (5) were unmanageable and hyperactive.

Of the children selected for inclusion in the test sample, 60 were eventually deemed untestable and 257 were found to be absent or unavailable during the week of testing at their center. In summary, then, 1,024 children were approached for testing, 964 were tested, and 60 (5.9 percent) were considered "untestable."

C. Principal Results

Exhibit 1 shows, for the different tests and program durations, the unweighted means, and the means obtained by making covariance adjustments to equate the samples for age, sex, race, size of town, and degree of poverty (p. 97). It is the adjusted means on which the main conclusion is based.

The principal result of the study was the absence of statistically reliable evidence of a treatment effect observed for the main samples of children tested, based on measures utilizing total test raw scores. No significant or systematic difference in mean scores associated with length of program was found. This conclusion rested on four criterion statistics:[1]

- Wilks' Lambda for multivariate covariance analyses.
- The F-ratios for individual instruments in the covariance analyses.

---

[1] Detailed descriptions of the data and data analyses can be found on pp. 42-124 and Appendixes B and F of the Final Report.

EXHIBIT 1 - UNADJUSTED AND ADJUSTED RAW SCORE MEANS[1]

Test	Program Duration	Unadjusted Mean	Adjusted Mean
Peabody Picture Vocabulary Test (PPVT)	Short	38.64	39.03
	Medium	41.73	39.57
	Long	38.07	40.34
Pre-School Inventory (PSI)	Short	46.50	46.48
	Medium	49.49	46.80
	Long	43.61	46.96
Vineland Social Maturity Scale (VSMS)	Short	55.49	55.58
	Medium	54.16	53.94
	Long	56.44	56.63
Behavior Inventory (BI)	Short	144.22	145.34
	Medium	139.98	138.74[2]
	Long	146.54	146.90

Notes: (1) Covariance adjustments were made for age, sex, race, population of community of the individual program, and degree of family poverty. The samples differed in composition with respect to these (and undoubtedly other) variables (see Exhibit 2).

(2) This mean was significantly lower than short and long Behavior Inventory means, at the .05 level of confidence (p. 98).

- The t-values (or confidence intervals for the β coefficient or weight for the exposure variable in multiple regression analyses.
- Proportions of test variance attributable to exposure time in multiple regression analyses.

This lack of evidence of a relationship between program duration and performance (or rating) of child applied whether the measures were considered jointly or (except in one case for the Behavior Inventory) individually.

If it is assumed that Head Start programs are capable of producing measurable cognitive, social, and emotional effects in children, what may have obscured the detection of such effects in this study? Before addressing this question, it should be understood that any discussion of effects involves only consequences measurable by the tests employed. The study was not concerned with medical or dental effects, for example, although these may well have been manifold. The meaning of an "effect," or at least the evidence for one, should also be clarified. In this study, one cannot speak of gain scores, that is, changes in a child's scores from the beginning to the end of the program.

As the experiment was designed, the short-term (S) group was the control or comparison group. The primary evidence for an effect would be a difference between the means of S and either or both medium-term (M) or long-term (L) samples for the dependent variables considered together or individually. There could, of course, be other indicators, but they may be less compelling, both in terms of power and operational significance. For example, there may be shifts in variability with no change in means; there may be consistent relationships between the means or medians of the samples (e.g., $L > M > S$) even though the differences are not significant by a parametric test; there may be weights in a multiple regression equation that are significantly different from zero, lending credence to the hypothesis of an effect associated with a treatment variable. However measured, whenever a treatment level distinction (S, M, L) is retained, an effect in this study was basically a mean score difference relative to the short-term program duration.

With these considerations as a framework, PRC examined five possible reasons or factors for the failure to detect the existence of any positive Head Start program effects in the samples studied. These factors were:

- Noncomparability of samples on essential uncontrolled variables (pp. 127-129).
- Inadequate sample sizes (pp. 129-131).
- Non-uniform treatments and effects among or within samples (pp. 131-134).
- Inappropriate instruments or tests (pp. 135-155).
- Effects of Head Start not immediately noticeable (p. 155).

No firm conclusions could be drawn about any of these possible obscuring factors. The noncomparability of samples (see Exhibit 2) raised many unresolved questions: Why was there a heavy loading of Spanish-speaking children in the short-term programs? Why were children in the medium-length programs generally white, rural, and older, while children in long-term programs were nonwhite, urban, and younger? Of course, the covariance analyses were an attempt to cope with these problems, but there may be important implications in the very fact that the distributions of characteristics of children varied so widely between the short-, medium-, and long-term programs.

A significant restraint in investigating other factors was the lack of information about local program objectives and goals (p. 155). However, because of the implications for program development, the discussion on non-uniformity of Head Start programs is presented here almost exactly as it appeared in the Final Report:

> Obviously, teachers, children, procedures, and facilities or environment varied from class to class as well as from CDC to CDC. Objectives and goals probably showed similar diversity. Consequently, whatever the specific nature of a Head Start center program, the "treatment" perforce varied at least from classroom to classroom. The results of the treatment undoubtedly varied from child to child. The issue is why the net result of the infinite specific and different treatments would be zero.
>
> The question is complex, and clearly unanswerable without recourse to specifics about treatments and effects--a specificity far beyond the available data in this study. Indeed, the specification and measurement of treatments in

EXHIBIT 2 - CHARACTERISTICS OF SHORT-, MEDIUM-, AND LONG-TERM ELIGIBLE CHILDREN

Characteristics	Short	Medium	Long	Total
Number	324	295	212	831
Mean Age in Months	61.37	61.85	57.67	60.45
Percentage of Boys	48.38	53.22	56.23	52.48
Percentage of Nonwhites	59.88	25.43	86.26	55.07
Mean Population of Community	138,000	47,000	1,400,000	190,000
Percentage of Poor	44.85	42.14	30.48	39.61
Percentage Tested in Spanish	16.63	2.07	3.51	7.42

child development and in education is still one of the most difficult problems in educational research.[1] Particularly in preschool programs, the treatment variables are difficult to define, except on a gross level, and more difficult to measure reliably.

There are undoubtedly interactions of treatments and subjects, however treatment is defined. Whether one thinks of variables as teachers or as center programs (or both), different levels (individually or in combination) very likely interact with pupil variables. Since the sample could not be defined to provide observations covering a known range of teacher-program-child variables, it could be argued that there were no apparent effects because there was an inadequate number of "positive" or "optimum" pupil-treatment observations.[2] Perhaps if classrooms had been used as the basic sampling unit, and all children had been tested in each sample classroom, there would have been sufficient numbers of specific teacher-program-pupil combinations at different levels so that an effect could have been observed. The net effect would presumably result from the occurrence of the differential effects. The theory and observation of a teacher- (and/or program-) child interaction is considered fundamental not only in education,[3] but in related enterprises such as psychotherapy.[4]

The problem here is not whether there were treatment-subject interactions, but what the net effect of interactions might be. For the design of this study, an effect measured by a difference, Δ, in S and L means (i.e., $\Delta = L - S$), could have one of three values: Δ is positive; Δ is negative; or Δ is, in effect, zero.

Let us assume, for purposes of discussion, that the experimental groups were comparable in starting level and that the two tests (PPVT and PSI) were each appropriate for measuring cognitive status or achievement level of the

---

[1] Gage, N. L., (ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963, passim

[2] Validity and sensitivity of the instruments are assumed for purposes of this argument.

[3] Gage, op. cit., passim

[4] See, for instance, Kiesler, Donald J., "Some Myths of Psychotherapy Research and the Search for a Paradigm," Psychological Bulletin, 1966, Vol. 65, pp. 110-136.

children. The measures themselves depended upon the responses of the subject to some partially known stimulus. The responses were binary (from the experimenter's point of view, a response is either right or wrong), and there was the usual confounding of cognition and motivation, which (in highly oversimplified and non-operational terms) is something like this:

Cognition	Motivation	
	Willing to Respond	Unwilling to Respond
Able to Respond	Correct Response	No Response or Wrong Response
Unable to Respond	Wrong Response or No Response	No Response

Here, the term "able" means "has the requisite knowledge and skills," while "willing" means "attempts to make best response possible according to perceived requirements (or rules) of the game (test)." As the relationships are depicted here, only the correct response has unambiguous meaning if we ignore the role of chance. In this study there were no specific criteria of motivation beyond the finding of the tester that the child was "testable" in the sense that he would stay in the situation and respond at all.

The effect of participation in a Head Start program could be in either or both realms and produce the same result. Conceptually, however, the action of a treatment may differ according to the area or realm of effect. That is, it is reasonable to conceive of cognitive effects as having a zero point and increasing in magnitude, complexity, scope, etc. It is more difficult to conceive of the cognitive effects as bipolar, with changes (losses) occurring as a result of participation in a program. On the other hand, motivational effects could easily be positive or negative, and could interact with cognitive effects in a variety of ways, uniformly or selectively (i.e., in terms of individuals or subgroups). However, regardless of what sort of treatment-pupil interaction is assumed, and regardless of what sorts of intereffect combinations are assumed, the absence of an observed treatment effect (i.e., difference in means) suggests that:

- There was no measurable effect in the cognitive or cognitive-motivational realms.[1]
- Positive effects in the cognitive area were nullified by opposing effects in the motivational area.
- Positive effects in the motivational area were nullified by negative effects in the cognitive area.

Another alternative, that positive effects in the motivational area were not accompanied by positive effects in the cognitive area, is tenable if, for example, it is assumed that short-, medium-, and long-term samples were systematically less developed in the cognitive and that the positive motivational effect simply maximized the use of otherwise unaffected knowledge and skills.

Any of the above hypotheses are possible, regardless of whether assumptions of uniformity or diversity of treatments and subjects are made. If one accepts the interaction point of view--that the major effects depend on the interaction of subjects and treatments--PRC's results suggest either (1) that there were too few optimum combinations to make a measurable difference (a scarcity that does not bode well from the point of view of matching teacher selection, training, program structure and content, teacher behavior, or whatever manipulable variables are considered the effective dimensions of a treatment with the appropriate pupil variables), or (2) that there were as many and as strong negative combinations or treatments and subjects as there were positive ones (pp. 131-134).

---

[1] This hypothesis in no way rules out the possibility of major effects occurring with all or many children early in their participation in a program (e.g., during the first 1 to 6 weeks).

Some evidence concerning diversity and/or lack of objectives, goals, and structure in general in some centers is discussed in Appendix E of the Final Report. It would appear that the issue of the role of structure in Head Start centers has not been resolved, and this may well reflect a philosophical split in the entire field of early childhood education. Comments such as the following occurred in testers' notes: "Classes were often mass chaos;" "Noise, constant interruptions, and general disorder;" "Inept and untrained teachers;" and "Teachers not aware of Head Start aims." There were, of course, other centers which featured hard-working, cooperative staff, but it is PRC's opinion

that a spelling out of specific goals and objectives would not only tend to provide much-needed direction for curriculum, facilities, and program development, but would also provide a framework for more effective staff selection, training, and supervision. If, as noted earlier, the overriding constraint for the overall study was "the lack of information about local program objectives and goals," it seems reasonable to assume that such a lack may also have been felt by the local Head Start staffs themselves.

With respect to the appropriateness of the test instruments, the Final Report, following an extensive discussion on the tests' validity, states:

> We have reviewed evidence available from our own data and offer the following comments as our opinions, based more on impressions than on cold analysis (p. 150):
>
> 1. For purposes of detecting general shifts in performance in a situation calling for use of receptive language skills and/or willingness and ability to operate according to the demands of authority (teacher, tester, etc.)--that is, to play the game--the PPVT seems to be fairly appropriate when differences in raw score means are used as the measures of effects. In our study, the PPVT was generally sensitive to variables to which it ought to be sensitive. To the extent that it simulates one form of situation or relationship which Head Start children, like others, will ineluctably encounter with increasing frequency and seriousness in their public school careers as the system presently works, it matters little whether the changes measured by it are cognitive, motivational, or both. Functionally, the result is the same. The challenge for Head Start is to find and clarify those procedures and techniques that maximize the development of effective cognitive skills, whether such techniques are directed at cognition, motivation, or both. It does not, in general, appear to have done so yet. We are not saying that other tests might not be equally or more appropriate. We are saying that, other things being equal (including, incidentally, administration costs), we think that the PPVT is reasonably appropriate for Head Start program evaluation purposes.
>
> 2. We think that the revised PSI is at least as appropriate for evaluation purposes as the PPVT. However, administration costs, including tester training, are higher. The PSI in many respects provides more information that is of operational significance than does the PPVT. Furthermore, the PSI appears a little more sensitive to

variables that one would expect it to be, and possibly a little less sensitive to confounding variables. The PSI was more sensitive to the urbanization measure than the PPVT (see Exhibit 45), although we did not examine individual items with respect to that variable. Somewhat more of the total PSI variance was accounted for by the independent variables considered in this study. Thus, the PSI is probably more sensitive to local conditions than the PPVT and as such makes a better instrument for local diagnostic purposes for children who were not extremely handicapped than does the PPVT. For our purposes, it seemed no more appropriate than the PPVT, and it was substantially more costly to administer and to train testers for than the PPVT. We have reservations about the order in which some items occur on the PSI (see Appendix D). We think that the present grouping of subtest items may enhance whatever test anxiety is inherent in the situation for low-income children. In some ways, the PSI seems constructed more to accommodate the academic standards of test specialists than to provide interpretable information about cognitive content or achievement. Nevertheless, it has at least a face validity for evaluation of Head Start programs that the PPVT lacks.

3. The Behavior Inventory remains an enigma. It told us little about effects (neither, of course, did other instruments). It told us relatively little about sensitivities to background or control variables. A number of teachers found many items ambiguous or hard to answer because of their multidimensionality. We found that the design of the form contributed substantially to the omission of responses by teachers and that a factor analysis of BI subtests gave us meaningless results. We found it difficult to imagine what criteria teachers used in making some ratings. Our overall opinion is that, before the BI is used further for diagnostic or evaluation purposes, systematic investigation and evaluation of it as an instrument should be undertaken. There is some evidence that it is grossly sensitive. However, there are too many uncertainties about what it really is measuring under conditions such as ours to recommend its further general use without more research and evaluation.

4. The Vineland Social Maturity Scale, as employed in this study, seems grossly appropriate, but not worth the cost of tester training and test administration. The negative beta coefficients for center size in the stepwise regression analysis, plus the negative beta for the amount of teacher's preschool experience, plus the significantly positive age betas (and correlations), suggest a tendency to "report" in terms of an age stereotype or not to report enough information. (Testers sometimes reported that teachers in larger centers seemed less knowledgeable

about or familiar with the children.) The overall average social quotients (SQ's) obtained in this study were about 100, as they should have been if this had been a typical group or if teachers had responded in terms of a typical stereotype. We found evidence of a tester-teacher interaction with the VSMS, and of a tester bias.

It is not clear that the VSMS would be sensitive to effects of Head Start treatment, even if the parent were the respondent. We think that the value of the VSMS in this study was to establish that the children in our sample probably were not conspicuously advanced or retarded for their ages in terms of the skills and abilities examined by the interview (pp. 151-153).

In regard to the possibility of latent effects, by definition of the study, they would not have been detected with a single end-of-program testing. The study states:

There is no question that longitudinal or cohort-tracking studies are vitally needed for major social programs and should, in the long run, provide more reliable and interpretable information than short-term studies such as ours. The dilemma, from a program point of view, is contained in the phrase 'in the long run.' Short-term studies have the most potential for influencing practical correction of discrepancies or inadequacies before procedures, operations, practices, attitudes, and so on become institutionalized. On the other hand, it appears that results from short-term studies are the more difficult to interpret both on technical and theoretical grounds. We are inclined to think that the trouble with short-term studies is just that; they are simply short-term studies. To the extent that they provide a base for continued observation, their value should be enhanced substantially (p. 155).

D. Conclusions

1. Subject to limitations in interpretation imposed by the design of the study, there was no statistically reliable evidence of a change in performance or rating of children in the major eligible samples on four test instruments which could be related to the length of a Head Start program or to the length of time that a child had attended a Head Start center. The conclusion also holds for various subsamples of children of similar age, sex, and race. However, the conclusion loses operational significance for subgroups as sample sizes decrease, since only very large changes can be assessed reliably when samples are very small.

The one significant variation in the test means between duration levels occurred with the Behavior Inventory. Various interpretations or explanations of the deviation were considered. PRC concluded that the variation was probably related to error of measurement and not to effect of the programs.

2. This conclusion does not vitiate the following hypotheses or possibilities concerning 1966 full-year Head Start programs:

- Children improved in many ways not measured by the tests, including health, nutritional status, and attitude toward schools and teachers.
- Children improved measurably on the dimensions measured by the tests or assessed by the rating or interview scales relatively early in their participation in programs.
- Parents, teachers, other staff members, and community organizations benefited from participation or involvement in the 1966 full-year programs.
- Beneficial effects of participation are latent and will become manifest after the children enter school.

None of these possibilities could be examined within the context of this study.

3. There were a number of factors which could have acted to obscure the observation and measurement of a Head Start treatment effect. Of these, the more significant methodologically appear to be:

- Lack of direct evidence that major experimental samples were comparable at the start of the programs.
- Some uncertainty concerning the validity and reliability of at least one of the measuring instruments.
- Lack of specificity of information about needs and goals associated with different programs.

The first is by far the most serious. As a result of the criteria used to identify the programs in the three duration levels studied, the distribution of centers was quite unlike any usual geographic distribution. Whether there were underlying selective factors differentially associated with the emergence of funded programs at different points in time during fiscal year 1966 is a matter of speculation.

4. The concept of an effect is complex and deserves close attention in the evaluation of large-scale programs aimed at changing behavior.

5. There is some evidence that the generally lower performance of Negro children relative to white children, especially as measured by the PPVT raw score, may be the result of a motivational rather than (or as well as) a cognitive factor. The situation that may cause the depressed scores is analogous to the situation and demands of the school classroom. If this is the case, it is certainly a condition which Head Start programs should be trying to correct.

6. Families of the children in the study were very similar in a number of characteristics in the different program levels. The characteristics of staff members and staff structures were generally similar for the three main samples, although some differences were noted. On a very gross basis of measurement, no significant relationship was observed (with one exception) between test scores and the amount of teachers' experience with preschoolers or with children from conditions of poverty.

In summary, this study conducted end-of-program tests of samples of children in 72 1966 full-year Head Start Child Development Centers representing programs of three main lengths or durations. The sample of programs ranged from 6 to 36 weeks in length at time of testing, and the average lengths of the three main samples were 12.4, 19.3, and 27.6 weeks. A sample of children in each center was tested with the Peabody Picture Vocabulary Test (PPVT), a test of general verbal ability, and the Caldwell-Soule Pre-School Inventory (PSI), which is designed to measure performance in several areas of social and cognitive achievement. Teachers were interviewed to provide ratings of the children on the Vineland Scale of Social Maturity (VSMS), and teachers also completed ratings of the children on the Operation Head Start Behavior Inventory (BI). The average test scores of the children in short-term centers (6 to 15 weeks) provide the basis of comparison for the examination of effects of Head Start programs on the children.

Statistical analyses of the results were undertaken in which a number of background variables such as age, sex, race, size of town, etc., were taken into consideration. As noted above, the overall result was that there was no significant indication of a general increase of scores with length of program.

The study did not examine the content or structure of the programs in the sample. Nor was any systematic attempt made to rate or evaluate the quality of the programs, personnel, or operations independently of the test scores. Consequently, PRC does not feel that the results mean either

that the programs accomplished nothing, or that many possible short- and long-term benefits to children, parents, and staff members did not occur. There are numerous variables involved in an enterprise as complex as Head Start, and the possible impacts and benefits are manifold. With respect to the functions, processes, or skills of children presumably assessed by the instruments used in this study, it appears that, overall, the gains to be expected with the longer programs exemplified by the 1966 full-year sample studied are small. The challenge, in the continued evaluation of programs such as Head Start, is to discover yet more precise, and at the same time comprehensive, means of depicting the true nature of the total array of benefits, and to translate such findings in further improvements in program design and operations (pp. 161-164).

COMMENTARY: "A Summary Report of A Study of the Full Year 1966 Head Start Programs"

John W. McDavid
Director
Research and Evaluation
Project Head Start

In conducting a study of the effects of Head Start attendance upon the behavior of preschool children as a function of the length of the Head Start program they attended, Planning Research Corporation was charged with the difficult task of evaluating a relatively specific hypothesis within an action setting which prohibited adequate isolation of the critical variable, program length, from a host of extraneous contaminants. Although data were collected carefully and analyzed exhaustively, the overall design of the study was subject to confounding effects which make it almost impossible to interpret finding of no differences in program impact as a function of program length.

The study was basically designed to evaluate the hypothesis that increased Head Start exposure (program length) would be associated with increased impact upon child development and behavior. However, there was no logical or empirical basis for expecting this association to be linear. In fact, it is inappropriate to conceive of the shorter programs (15 weeks or less) as the proper baseline for comparison against the longer programs (as stated on page 9), since a number of evaluative studies have established that significant changes in children's behavior occur within the very brief (6 to 8 week) summer Head Start programs. Thus, it was possible that highly important modifications of the child's behavior may have been produced even earlier than the shortest interval evaluated in this study, as noted on pages 14 and 18 of this summary report.

The fact that this study was designed as a post-treatment comparison of three unmatched groups of Head Start children, affords no assurance that the three groups entered the Head Start experience at comparable levels of performance. It was, unfortunately, not feasible during the 1965-66 to conduct a pre/post comparison study. Consequently, proper baseline performance data for each of the three groups were simply not available.

The lack of specificity of focus and direction in the 1965-66 Full Year programs described on pages 14 and 15 of the Summary Report may be in part attributable to the fact that these programs were the pilot advance guard for later evolution of the Full Year Head Start program. All were operating for the first time with newly employed and generally inexperienced staff, and their program objectives and styles of operation had not yet achieved stability. This vagueness and lack of specificity would certainly be expected to contribute to random error variation among programs (as discussed on page 10 of the Summary Report). However, some variability among programs may reasonably be expected to continue in Head Start as a nation-wide program.

Head Start operates within a general philosophy of local autonomy which authorizes and encourages local programs to determine their own style of operation according to the needs of local populations, within broadly defined guideline boundaries defining good preschool practice and the overall goals of Project Head Start.

The most useful kind of information gained from the PRC study is that reported on pages 10 and 11 of the Summary Report: that differing characteristics of the population served were associated with programs of varying length during 1965-66. Furthermore, the original detailed report from which this Summary Report is derived outlined certain anecdotal evidence that program style and operation also differed in programs of different durations.

It would appear, for example, that middle-sized communities with relatively large concentrations of Spanish-speaking children elected during 1965-66 to conduct relatively short length programs, while large cities elected to conduct programs of longer duration for relatively younger non-White children. However, even such extraneous variables as the logistics of application and funding may be operative in determining these relationships. These systematic differences confounded with program length contribute unspecified error within the design of the study reported here, hindering the attempt to isolate the effects of program length variation as such.

In effect, then, it would seem that this study of program length was conducted prematurely. It would have been more appropriate to design first an elaborate national evaluation of Head Start to ascertain the major critical variables associated with program style and population characteristics which determine Head Start's impact upon the child's performance, and then subsequently to attempt to evaluate the effects of program length with proper controls to isolate other critical variables apart from program length. In fact, such a strategy is now feasible, since the overall national evaluation of Head Start Full Year programs for 1966-67 and 1967-68 has been designed to produce the empirical data base to permit such refinement of a future design for evaluating the effects of program length.